PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/30 — A1

(11) International Publication Number: WO 00/54185

(43) International Publication Date: 14 September 2000 (14.09.00)

(21) International Application Number: PCT/US00/06072

(22) International Filing Date: 8 March 2000 (08.03.00)

(30) Priority Data:
09/264,299 8 March 1999 (08.03.99) US

(71) Applicant: THE PROCTER & GAMBLE COMPANY [US/US]; One Procter & Gamble Plaza, Cincinnati, OH 45202 (US).

(72) Inventors: KIRKPATRICK, James, Frederick, Jr.; 12 High Street, Milford, OH 45150 (US). TOOGOOD, Kevin, Charles; 7455 Huckleberry Lane, Cincinnati, OH 45242 (US).

(74) Agents: REED, T., David et al.; The Procter & Gamble Company, 5299 Spring Grove Avenue, Cincinnati, OH 45217–1087 (US).

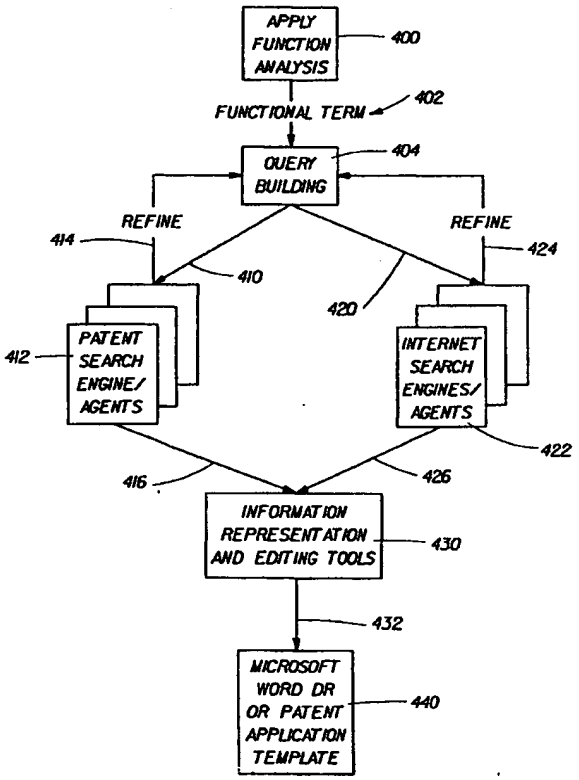
(81) Designated States: AE, AL, AM, AT, AT (Utility model), AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, CZ (Utility model), DE, DE (Utility model), DK, DK (Utility model), DM, EE, EE (Utility model), ES, FI, FI (Utility model), GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (Utility model), SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR BUILDING A USER–DEFINED TECHNICAL THESAURUS USING ON–LINE DATABASES

(57) Abstract

An on–line data collection system is provided that uses a network to access multiple databases over the Internet (Figure 5). The user designates an initial search term and performs a preliminary search (or "initial query") to obtain preliminary results that rank documents containing the initial search term (402). Using this information, the system can show words or phrases equivalent to the initial search term, and each of the search terms also can be associated with an image that is chosen by the user. The system iteratively searches databases containing "electronic" documents using all desired terms completely under the control of the user, and as equivalent terms or new terms are discovered within certain documents, the user can refine the search query (404) to finally arrive at a "crafted query". This crafted query will produce a set of documents that are grouped by certain criteria, and these various criteria can then be linked so that more meaningful relationships between certain documents can be easily discerned by the user. The end result is a very useful interactive database searching tool that allows the user to quickly find the documents that are most relevant to that user.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHOD AND APPARATUS FOR BUILDING A USER-DEFINED TECHNICAL THESAURUS USING ON-LINE DATABASES

TECHNICAL FIELD

The present invention relates generally to on-line data collection and transmission equipment and is particularly directed to computer networks of the type which access databases over the Internet. The invention is specifically disclosed as an interactive data gathering system that builds technical thesauri specifically tailored by a human user.

BACKGROUND OF THE INVENTION

The Internet has provided free access to many technical documents, including United States Patents, to every engineer and scientist that is equipped with a computer, a modem, and a browser. However, it is impossible for an individual human mind to encompass this vast expanse of data. Knowledge of the boundaries of existing technology is nevertheless a pre-condition for cost-effective research and development as opposed to blind efforts that may recreate what already is known in the prior art.

Software tools are already being developed to automate data gathering and presentation of technical documents, including a United States Patent No. 5,621,910 (by Unger) which discloses a database that is used to determine the meaning of scientific or technical documents and to assign these documents to particular categories. The Unger system has a primary thrust of identifying trends or discontinuities in the research efforts of competitive companies. Once categories are determined, the Unger system can display results on a spreadsheet or in a graphical format, such as a bar chart.

In US 5,652,829 (by Hong), a "feature merit generator" is disclosed that classifies a document and correlating its context by determining a "classification merit factor" as an indicia of a vector of feature variables. This merit factor determination occurs by comparing the feature's "obligational dependency" to a class discrimination in a context of

corresponding feature variables in at least one example which generates a feature variable ranking for the given class of interest.

US 5,467,425 (by Lau) discloses an n-gram language modeler which reduces the memory storage requirements and convergence time for language modeling systems. The Lau invention aligns each n-gram with one of the "n" number of non-intersecting classes. A count is determined for each n-gram representing the number of times each n-gram occurs in the training data. A word predicting language modeler finds the probability that a word occurs given the occurrence of two previous words.

US 5,418,951 (by Damashek) discloses a method of identifying, retrieving, or sorting documents by language or by topic using an n-gram array for each document in a database. Each unidentified document is parsed into n-grams, a weight is assigned to each n-gram, the commonality is removed from the n-grams, and each unidentified document is compared to each database document. The unidentified document is then scored against each database document for similarity, and based upon the similarity score, the document is sorted with respect to language or a topic.

US 5,576,954 (by Driscoll) discloses a method for determining text relevancy of documents being retrieved by search queries. The Driscoll system helps a user intelligently and rapidly locate information found in large textual databases by determining common meanings between each word in the query and in the document being retrieved. Weights are calculated, multiplied, and added to create a "real number" similarity coefficient for the document. The documents are then sorted in sequential order according to their real number value from the largest to the smallest. A second embodiment can route documents based upon topics or headings, also referred to as "filtering." The importance of each word in both the topics and the documents are calculated, and the real number (similarity coefficient) for each document is determined. Each document is then routed according to the similarity coefficient.

US 4,823,306 (by Barbic) discloses a method of text searching of documents. Starting with a word query, a set of equivalent words are defined for each query word along

with a corresponding word equivalences value that is assigned to each equivalent word. Target sequences of words in a library document that match the sequence of query words are located according to a set of "matching criteria." The similarity value of each target sequence is evaluated as a function of the corresponding equivalence values of the words included in the target sequence. A relevance factor is then obtained for each library document based upon the similarity values of its target sequences.

US 5,555,354 (by Strasnick) discloses a method for navigating within a three dimensional graphic display space, and manipulating information represented by objects in the display space. Data objects represented by graphic objects are arranged into a navigable landscape representing the relations of the underlying data. The graphic objects comprise columns, pedestals and disks, which respectfully represent datablocks, cells, and comparative values. The ground plane represents a threshold value, upon which the pedestals rest. Data attributes may be represented by visual, textural, or audible characteristics of the display. The user can interact with the data to make changes in the underlying data or its representation within the display space. Less detail is displayed as the user navigates away from objects within the display space. Objects change from three-dimensional to two-dimensional, then to line segments as the user moves away from the objects.

The conventional data searching systems noted above are not designed for any type of interactive session during the creative process, but instead act as a batch input. It would be an improvement in searching text documents for a computer program to assist the user in creating more intelligent queries, especially as an interactive procedure that makes it easy for the user to sort out important words and phrases from non-important words and phrases.

SUMMARY OF THE INVENTION

Accordingly, it is an advantage of the present invention to provide a computerized system that interactively enables a user to build a technical thesaurus using on-line databases. It is another advantage of the present invention to provide a computerized system that iteratively prompts a user to interactively modify an "initial query" into a "crafted query" while searching at least one on-line database. It is a further advantage of the

present invention to provide a computerized database searching system that searches on-line technical documents based upon an interactively-created query that uses both exact search terms and equivalent search terms. It is yet another advantage of the present invention to provide a computerized database searching system in which the search terms to create queries are associated with an illustration or image for ease of visual reference when later accessed by the user. It is yet a further advantage of the present invention to provide a computerized database searching system that allows a user to choose terminology used to create a crafted query, in which technical documents can be searched by various fields that can be linked to common terminology by one or more linking criteria.

Additional advantages and other novel features of the invention will be set forth in part in the description that follows and in part will become apparent to those skilled in the art upon examination of the following or may be learned with the practice of the invention.

To achieve the foregoing and other advantages, and in accordance with one aspect of the present invention, an improved on-line data collection system is provided that uses a network to access multiple databases over the Internet. By use of the Internet (or some other local or wide area network), a human user may access "electronic" technical databases, including those that contain patent documents. The user begins by designating an initial search term and performing at least one preliminary search (or "initial query") to obtain preliminary results that can rank technical documents which contain the initial search term. Using this information, the computerized system can show expressions (i.e., words or phrases) equivalent to the initial search term, and these equivalent terms can also be selected by the user for association with the initial search term, or can be used to create entirely new search terms, if desired.

Multiple search terms (with or without their equivalents) can be used to search through one or more databases containing technical documents. Each of the search terms also can be associated with some type of illustration or image that is chosen by the user. Such visual images will be more easily recognizable by the user in many circumstances, and

preferably will be located on the monitor screen in close proximity to the displaying of the search term.

The computerized system can iteratively search the database or databases containing the technical documents using the initial search term as well as all desired new terms (with or without their equivalents), completely under the control of the user. As equivalent terms or new terms are discovered within certain documents, the user can begin to refine his or her search query. As part of this refinement, the user may discover certain terms or phrases that are not desirable with respect to being included in a particular search query. The user can use "NOT" Boolean connectors to preclude finding any documents that contain certain terms that are not desired as part of the query. This is all part of the interactive aspect of the present invention.

After a sufficient number of iterations of preliminary searches have been interactively performed, the user will eventually arrive at a final "crafted query," which is then used to again search through the same or other databases of technical documents. This crafted query will produce a list or set of documents that are grouped by certain criteria, such as documents published within a certain range of dates. The various criteria can then be linked, if desired by the user, so that more meaningful relationships between certain documents can be easily discerned by the user. The linking criteria can also be on a sliding scale, as determined solely by the user, which allows the user to choose certain linking criteria that are more important than others. The information can be presented in various manners, including a ranking of the frequency of occurrence of a particular search term per document, or the frequency of occurrence in all documents of a particular search term.

The various fields that appear on the monitor screen can be made to be interactive with one another by clicking and dragging information from one field to the next. In this manner, the creation of the crafted query can be done very efficiently, while providing appropriate user interaction with the computerized system. The user can create personalized information as part of his or her crafted query, including a list of words that are unique to a particular user. In this manner, various technical documents (including patents) can be stored in the user's personal thesaurus, and this thesaurus can be later built upon for linking

with other thesauri, either created by the same user or by a different user who is working in the same computerized system from time to time.

The search engines used by the user either can be manually operated (in real time while the user is using the Internet or another network), or can be used with an automatic search agent that will remain on-line up to 24 hours per day to search through large databases containing many documents. Such large databases exist, especially with respect to patent documents, on the Internet. The final result is a list of useful documents that have been "found" by the computerized system using the user's personal crafted query, which provides the documents that are most useful to the user for his or her particular subject matter for the time being. Other crafted queries can be used to gather related documents, and linking criteria can be used to compare and contrast documents from more than one crafted query search. Each crafted query search can be used to create an individual thesaurus, or many crafted queries can be used to create a single thesaurus, which is completely dependent upon the desired database structure by the user. The end result is a very useful interactive database searching tool that allows the user to quickly find the technical documents that are the most relevant to that user.

Still other advantages of the present invention will become apparent to those skilled in this art from the following description and drawings wherein there is described and shown a preferred embodiment of this invention in one of the best modes contemplated for carrying out the invention. As will be realized, the invention is capable of other different embodiments, and its several details are capable of modification in various, obvious aspects all without departing from the invention. Accordingly, the drawings and descriptions will be regarded as illustrative in nature and not as restrictive.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings incorporated in and forming a part of the specification illustrate several aspects of the present invention, and together with the description and claims serve to explain the principles of the invention. In the drawings:

Figure 1 is a block diagram of the major components of a data collecting and thesaurus building system, as constructed according to the principles of the present invention.

Figure 2 is an example display used in setting up a search term, according to the principles of the present invention.

Figure 3 is an example display similar to Figure 2, with the use of further search terms.

Figure 4 is an example display similar to Figures 2 and 3, further indicating the building of a search query.

Figure 5 is a flow chart of the overall invention's method in broad terms from the first step taken by the user to the final step of displaying and integrating the result of a search.

Figure 6 is a flow chart of the detailed steps used in crafting queries, as determined by a human user.

Figure 7 is a flow chart showing the major steps utilized by the present invention after a query has been crafted to link the results into one or more databases.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Reference will now be made in detail to the present preferred embodiment of the invention, an example of which is illustrated in the accompanying drawings, wherein like numerals indicate the same elements throughout the views.

Referring now to the drawings, Figure 1 shows a personal computer or workstation generally designated by the reference numeral 10, which is electrically connected to a video monitor 12 and a printer 14. Video monitor 12 is controlled by a video controller 20 that is contained within the computer 10, and is connected to the monitor by a video cable 22.

Computer 10 is controlled by some type of central processing unit (CPU) 30, which typically comprises a microprocessor. CPU 30 controls virtually everything that occurs within the computer 10, and communicates with the other electronic components via a data/address/control bus 28. CPU 30 accesses memory circuits 32, such as random access memory (RAM) and read only memory (ROM). In the present invention, it would be desirable to have a bulk memory storage device 34, such as a large hard disk drive. Depending upon the amount of data to be stored, a read/write compact disc drive might be even more useful than a hard disk drive for use as the bulk memory storage device 34.

Assuming a printer is desirable, the printer 14 is connected by a printer cable 26 into a communications port 24 of the computer 10. The a printer could be connected through a network, rather than having a direct connection as shown on Figure 1. The communications ports 24 will preferably also include a modem that can connect to an outside communication link 52, such as a telephone line. This communication link 52 is connected to a network service provider 50, which is sometimes referred to as an "ISP" for "Internet Service Provider."

Computer 10 will require some type of keyboard 42 and some type of cursor pointing device, such as a mouse 46. Keyboard 42 is connected via a keyboard cable 44 into an input/output interface circuit 40. Mouse 46 is connected through a mouse cable 48 into another port of the input/output interface 40.

As is commonly known, the Internet is a world-wide network of computers that are accessible by virtually anybody having a computer with a modem. On Figure 1, the Internet is depicted by the reference numeral 60. The user of computer 10 accesses the Internet 60 using his or her network service provider 50. Once connected into the Internet, this user can access information at other web sites.

On Figure 1, particular web sites that contain patent information are depicted, such as the United States Patent and Trademark Office (PTO) 72, the European Patent Office (EPO) 74, the Japanese Patent Office (JPO) 76, and the "DIPS" world patent database 78.

Each of these government or commercial institutions must also have some type of Internet service provider, such as an ISP 62 for the PTO, ISP 64 for the EPO, ISP 66 for the JPO, and ISP 68 for the DIPS database 78.

Of course, it is possible for users on the Internet to access other sources of technical documents, even including other sources of patents. For example, IBM has a web site of United States patents that are accessible at "http://patent.womplex.ibm.com." This IBM patent database can be used to search for various patents containing particular words that are chosen by the user. It can also be searched according to other classifications, such as inventor's name or assignee, also chosen by the user.

If other technical documents are of interest to the user, then certainly there are thousands of Internet web sites that contain such documents which can be easily accessed over the Internet 60 by the user of computer 10. These documents may not be readily searchable by any type of searching tool provided by the owner of the Internet web site, however, the user of computer 10 can provide his or her own searching mechanisms, if desirable. One brute force method for searching a document is to download the document onto the bulk memory storage device 34, and later inspect the file that holds that document by whatever displaying and searching tools the user may have available at computer 10.

Figure 2 illustrates a "set-up screen" designated by the reference numeral 100, which can be viewed on video monitor 12 by a human user. In the preferred mode of operation, set-up screen 100 will be interactive with the human user, and also with an on-line database over the Internet or other type of network.

Starting with an initial term designated at the reference numeral 112 as "Term 1," the user can not only choose a particular word as "Term 1," but can also choose other equivalent words. On Figure 2, a column 110 includes Term 1 at 112, and also includes three different equivalent words or phrases at 114, as indicated by the terminology "Equivalent 1," "Equivalent 2," "Equivalent . . . ." These words that are chosen by the user as equivalents are not necessarily chosen right at the beginning of the procedure for creating a search query. Instead, the user may choose Term 1, and actually go into a patent database,

for example, to see what types of patent documents are produced by such a search query. From this initial search, the user may learn that other words or phrases are more or less equivalent to that initially chosen as Term 1. If these other equivalent words or phrases are considered desirable by the user, then these equivalent words and phrases can be added into the column 110 on Figure 2, at the locations designated by 114.

Column 110 also includes an area at the reference numeral 130 where a drawing or other image (such as a photograph or computer generated graphics) can be "linked" to Term 1. Until a particular image or drawing is chosen, the preferred embodiment displays a phrase "LINK TO?" This phrase is merely a reminder to the user that he or she may link an image or drawing with Term 1, but this is not a requirement in the preferred embodiment. When the search results are presented to the user, as will be discussed below, the image or drawing at 130 will follow the presentation of the search results on the screen of video monitor 12. In this manner, an easily recognizable pictorial image can be forever linked to Term 1, much as a logo when associated with goods in the trademark sense. Therefore, when many different terms are being displayed on the same screen, the visual images will be more easily recognizable by the user in many circumstances than the actual words that make up Term 1.

Other terms can be placed in different columns, which on Figure 2 are blank because they have not yet been set up by the user. A column 120 placed in the upper right-corner of display 100 is used to view the preliminary results of a search. Before a query is launched, one or more of the terms must be placed into the "Found" column 120. On Figure 2, Term 1 has been placed into this column, at the reference numeral 122. This term can be brought over from column 110 by clicking and dragging, or by clicking and choosing. Once the query for Term 1 is launched, then the preliminary results will be displayed in column 120 (as seen on the later figures).

When more than one term is placed into the "Found" column 120, then the arrows 102 and 104 can be used to choose one of the terms. In addition, a moveable scroll bar 106 can be used to view any particular terms that are displayed in the Found column, especially

at times when more terms reside in the Found column than there is area to display at one time.

A set of Boolean operators are found in the lower left-hand corner of display 100 at the column designated by the reference numeral 140. These Boolean operators are used to create sophisticated search queries, and this aspect of the present invention will be discussed in greater detail hereinbelow.

On Figure 3, four different columns contain various search terms of a display 200 that can be viewed on the video monitor 12. The four columns that contain search terms are designated by the reference numerals 210, 220, 230, and 240. Each of these columns has a main "term" that will be entered at the top of each individual column, at the reference numerals 212, 222, 232, and 242. Each of these individual terms can have equivalents, as illustrated within each of the columns, at the reference numerals 214, 224, 234, and 244.

The "Found" column at 250 is depicted as having eight different search terms at this point in the search query procedure. If patents, for example, are the types of documents that are being searched, then the present invention is able to limit a search of an on-line database by choosing certain attributes that will "channel" the field of the search. For example, the patent database being inspected could be accessed by looking for attributes such as the name of the inventor, name of assignee, by patents issuing within a certain span of dates, or by the United States Class in the PTO records.

Assuming a particular preliminary search has already been performed using the eight terms displayed in column 250 on Figure 3, a sub-column at the reference numeral 260 will show the number of "hits" for each of these terms based upon the portion of the patent database that was inspected by this particular search query. For example, the "Term 1" line at the reference numeral 251 on Figure 3 shows a number of hits as being "233/20." This represents the fact that the Term 1 query was found in 20 different documents, and was found a total number of 233 times within those 20 documents.

In a similar fashion, "Term 2" at 252 was found 154 times in 15 documents, "Term 3" at 253 was found 127 times in 10 different documents, "Term 4" at 254 was found 100 times in 10 different documents, "Term 5" at 255 was found 86 times in ten different documents, and "Term 6" at 256 was found 23 times in eight separate documents. As according to the preferred embodiment, the term that was found the most often is typically listed first in the "Found" column 250.

Since there is not enough space on display 200 to list in detail all eight of the terms being searched, the present invention allows the user to choose from this list the terms that are of the most interest and to have them placed into the columns 210, 220, 230, or 240. In the illustrated example on Figure 3, the "Term 4" was not chosen by the user, but instead "Term 5" was chosen to be placed in column 240. Of course, the user could click on "Term 4" at 254 in the "Found" column 250 and drag that Term 4 over to one of the other columns 210, 220, 230, or 240, if the user desired to see Term 4 in greater detail. The user thus can easily inspect the exact terminology being used for the equivalents of each of the terms that are displayed in the found column 250.

Some type of illustration may also be associated with each of the terms, along the lines as related above with respect to Figure 2. For example, Term 1 could have had a drawing or other type of image linked with itself at 216. The same is true for Term 2 at 226. Term 3 at column 230 has been associated with a particular drawing, as illustrated at the reference numeral 236. The same is true for Term 5 in column 240, which has a drawing at 246 linked to Term 5. Finally, one of the terms listed in the Found column 250 can be selected by the user, and its associated illustration or drawing will automatically appear within the small window at 262. This is made possible by a "live" scrolling of the terms within the window that make up the Found column 250. The UP button 202 and DOWN button 204 can be used to select a particular term within the Found column 250. Furthermore, a scroll bar at 206 can be used to view other terms that may temporarily be not visible within the window at 250.

The use of the terminology "window" is appropriate in the present invention, because its preferred operating system is MICROSOFT WINDOWS 98™. By use the

WINDOWS 98 operating system, the user can view two different types of displays simultaneously on two different video monitors. For example, the user could display one of the set-up screens (such as the screen 200 on Figure 3) on the first video monitor, and could perhaps show a more detailed set of results of a search query on the other video monitor.

The set-up display 200 also includes the Boolean operators in the lower left-hand corner in a column 270. One of the operators, at the reference numeral 272, can be used to link an image to one of the terms that are displayed in the top-half of the display 200. By use of the "Link Image" button at 272, the user can always change the illustration (or image) that is associated with a particular one of the terms.

Figure 4 illustrates another set-up screen 300 used in creating a visual query while building a technical thesaurus according to the present invention. Four different terms 312, 322, 332, and 342 have been selected for the columns 310, 320, 330, and 340, respectively, by the user highlighting those terms at 351, 352, 353, and 355 in the "Found" column 350. On Figure 4, for example, Terms 4 and 6 at 354, and 356, respectively have not been selected by the user at this time. Scroll buttons 302 and 304 and a scroll bar 306 are used to view the different terms that have been located in the "Found" column 350.

As in the previously described set-up screens on Figures 2 and 3, "Term 1" 312 can have equivalent words or phrases associated therewith, as illustrated at 314. Similarly, "Term 2" 322 can have equivalent words and phrases at 324, as well as "Term 3" at 334 and "Term 5" at 344. Each of these columns can also have an illustration linked therewith, such as at 336 for column 330, 346 for column 340, and 362 for the term that is presently selected in the live scrolling window of the "Found" column 350. Columns 310 and 320 have "Linked To" areas 316 and 326, respectively, which can be used to illustrate a drawing or other type of image, if desired by the user.

The "Link Image" button 372 is available for use by the user to choose a particular drawing or image to be associated with one of the terms. A series of Boolean connectors is provided at a column 370 in the lower left-hand corner of the set-up screen 300. These

connectors include an AND button 374, and OR button 375, a NOT button 376, a left parenthesis button 377, and a right parenthesis button 378.

The Boolean connectors 374-378 are used when a sophisticated search query is to be built by the user at 380. On Figure 4, an example search has been provided which includes "Term 1" and "Term 2," as well as their equivalents. Term 1 at 382 is combined with its equivalent terms 384 by the restrictive Boolean connector AND at 390. Term 2 at 386 is combined with its equivalent terms 388 by the restrictive Boolean connector AND at 394. Both of these combinations are combined themselves by the expanding Boolean connector OR at 392.

Other search terms with or without their equivalents can be added into the search query that is being built within the window 380 in any combinations of Boolean connectors desired by the user. One major improvement provided by the present invention is the ability to exclude a certain term with its equivalents, if desirable, by using the NOT Boolean connector. For example, the query being built in window 380 on Figure 4 could also have "Term 5" added, while being preceded by the NOT connector. In other words, the search would then comprise Term 1 and its equivalents, OR Term 2 and its equivalents, but NOT Term 5. Furthermore, this NOT restriction to the search could be for Term 5 and its equivalents. In this manner, the user may build a very intelligent search based upon words and phrases that have been previously discovered by accessing a prior art patent database which discovered a particular group of words (e.g., Term 5) that are not desirable to be contained within the user's search query. In this manner, the user may refine his or her search to provide a very meaningful result, with minimum effort while accessing a patent database.

The potential interaction between the computer 10 and the user can lead to a very powerful searching tool that is very selective, but also at the same time can be aimed at a large number of various search terms while being quite easy to use. For example, if one of the terms chosen by the user in the top half of the set-up screen 300 already resides in one of the thesauri of the bulk memory storage device 34, then that term and all of its equivalents will be called up from the particular thesaurus that contains that term, and all of the

equivalent terms can be embedded in the search. Naturally, the user can decide how best to use equivalent terms, and may decide to either include such equivalent terms in or to exclude them from a particular search query. Furthermore, a particular word or phrase can be an equivalent to more than one of the terms along the top of the set-up screen 300. In fact, one of the equivalent words or phrases could become its own search term, if desirable.

If the operating system used is WINDOWS 98, then navigation through the set-up screen will be quite easy for a person that has a rudimentary knowledge of windows-type computer programs. The various terms and equivalents can be chosen by clicking and dragging or clicking and choosing, and can be "dropped" into virtually any other of the areas on the display. The main purpose, of course, is to create a search query in the window 380 that is useful in searching through a large number of patent documents or other types of technical documents. It will be understood that other arrangements of displays could be used as set-up screens without departing from the principles of the present invention. Furthermore, different types or numbers of Boolean connectors and different numbers or arrangements of search term windows and "Found" windows could be used without departing from the principles of the present invention.

An overview of the broad method steps used in the present invention is provided in Figure 5. Starting at an initial step 400, a research scientist or engineer would be a typical user who would have a need to create new technology based upon a functional analysis of existing technology. This user would provide one or more functional terms (at reference numeral 402) to a query building step 404. The query building step 404 is an interactive procedure, as described hereinabove in connection with the set-up displays on Figures 2-4.

The query building step 404 can also receive functional terms from input sources other than a functional analysis step, such as the step 400. This would strictly be up to the human user, and may depend upon whether or not the user is a technical person or has a non-technical background.

The query building step 404 is very interactive with input both from the human user and from various types of databases that can be accessed over a network. For example, one

or more patent search engines or search agents, generally represented by the reference numeral 412, can be accessed over a network link 410. The results of accessing the patent search engine/agents are provided over a return path 414, and these results are iteratively refined by the query building step 404 in the preferred mode of the present invention.

Another form of information can be found using Internet search engines or Internet agents, which are generally designated by the reference numeral 422. These Internet engines/agents are accessed over a network pathway 420, and their results are provided over a pathway 424. These results over the pathway 424 are iteratively refined by the query building step 404.

After the information has been refined by the query building step 404, the results sought by the user are provided from the patent search engine/agents over a pathway 416, or from the Internet search engines/agents over a pathway 426. The information from these pathways 416 and 426 is provided to an application module 430 that presents information to the user and allows the user to use editing tools to distill and refine that information. The final information derived from this step 430 is provided via a pathway 432 to some type of MICROSOFT WORD file, or perhaps to a patent application template, at a step 440.

The use of MICROSOFT WORD file would be useful in creating inventive development records, or other types of similar documentation describing inventive or technical documents. Of course, other types of word processors could be used than the MICROSOFT WORD computer program.

A patent application template would preferably include blocks of information to receive prior art background information, for example, and also perhaps information that could be used to highlight a new invention, even with proposed concepts that could be claimed. It could be used to begin generating a U.S. Patent and Trademark Office disclosure document, or even a provisional patent application.

Figure 6 is a flow chart of the detailed steps used in the interactive query editing and building steps of the present invention. Starting at a step 500, the human user begins to

create a visual query using the editor that is illustrated on Figures 2-4. At this point in the procedure, the user would be viewing a display screen such as that illustrated on Figure 2, and would be working only with text at this point.

After one or more initial search terms have been chosen by the user, the present invention searches one or more network-linked databases using some type of search engine. For example, a "META" search engine at a function step 502 can be used to connect to the Internet and act as a automatic search agent. An alternative is to use a site specific agent at a function step 504, such as the IBM patent database that is found on the Internet at " http://patent.womplex.ibm.com." As a further alternative, a search engine agent could be used to access the United States Patent and Trademark Office records that are available over the Internet at a function step 506. Naturally, other types of search engine agents and other types of databases can be used when using this portion of the procedure of the present invention. At this point in the procedure, it is strictly up to the user as to which type of search engines and databases that are chosen to be searched, and the user is not restricted to using the Internet, but could link into any type of database available to his or her computer 10.

The next step in the procedure of the present invention is illustrated as a function step 510 where the user parses text while eliminating non-descriptive language from the search terms and stemming of descriptive language, thereby keeping the important words and/or phrases. This step 510 is reached from any of the searching agents, such as agents 502, 504, and 506, and also can be reached by entry of a patent document or other type of existing technical document that may be available directly to the user, from a step 512, on Figure 6.

The procedure of the present invention now reaches a step 514 where the stem words are ranked and presented for user selection. Regardless of the source of the document, the present invention will parse through it, and after its analysis will then rank the documents in situations where more than one document is being analyzed. In this circumstance, the video monitor 12 will display a screen such as the screen 200 that is illustrated on Figure 3. On Figure 3, the "Found" column 250 shows a ranking of the most

popular terms, at least as far as how many times a particular term was found and in how many documents.

This information can be presented in more than one way, such as the frequency of occurrence of a search term per document, or the frequency of occurrence for all documents.

At step 514, the user can click and drag words or phrases to add or delete from an existing query. These words or phrases can be manually entered, or the user may take existing text from a patent document or other technical document and highlight phrases or words in that text and link them to equivalent terms in the technical thesaurus.

The final result of step 514 is for the user to have created a complete query, although this will typically not occur the first time through the procedure of the present invention. Therefore, this query building process is an iterative process, and the logic flow is directed to a decision step 516 that asks whether or not the query is complete. If the answer is NO, then the logic flow is directed back to the visual query editor/builder step 500. If the answer is YES, then the logic flow is directed to a function 530 which represents a number of different steps that are illustrated on Figure 7.

If the query is incomplete at the decision step 516, then the logic flow not only is directly back to the step 500, but a function step 520 is used to link synonyms and images. Information gleaned from step 520 will be shared with the visual query editor/builder step 500. As part of this linking routine 520, data is stored on the bulk memory storage device 34 at a step 522 to create one or more thesauri with linked images. The steps 520 and 522 allow the user to choose certain close synonyms, if desirable, and to create a list of words that are unique to that particular user.

On Figure 7 the flow chart begins with a step 540 which is arrived at only after the complete "crafted query" has been obtained. An example of this type of query is shown in the window 380 on Figure 4. This is the type of query that the present invention is useful in creating, since it allows the user to select the exact query terms that will be used in the search, while also selecting precise terms that will be inhibited during the search.

Furthermore, any of these terms can include equivalent terms, such as the equivalent terms 314 that are associated with "Term 1" 312 on column 310 of Figure 4.

After the crafted query is available at step 540, various database search engines and agents can be utilized for the crafted query search. These can be the exact same search agents as used before on Figure 6, for example, at 502, 504, and 506. On Figure 7, these same agents are respectively designated by the reference numerals 542, 544, and 546. Naturally, many other search agents and databases can be accessed, which provides an information source that is virtually limitless within the resources of the user. Of course, if the Internet is used, then the user will essentially be limited only by the data communications rate and the number of hours in the day that he or she is accessing one or more Internet databases while looking for technical documents of interest. On the other hand, if the user has an automatic search engine resident on his or her computer 10, then the Internet database search could go on literally 24 hours per day, with very little guidance from the user during a portion of these time periods.

As the search engines find documents that include the crafted query terminology, both image and text data can be linked to Internet pages by an in-memory operation at a function step 550. For example, United States patents can be searched by various fields, such as the name of the inventor or the name of the assignee, as well as patent drawing images that are available at certain Internet patent servers. Any of these fields for images can be linked to common terms that may appear in the patent fields, or the search query terms from the crafted query step 540.

The linking criteria is provided by a function at a step 552, in which the linking criteria establishes what downloaded data is linked to a file and what is discarded. Linking criteria can include attributes, such as: (1) file size, such as a maximum size per document, (2) image size, (3) image proportions, (4) image type, such as "gif", "jpeg", etc., which with knowing the image proportion can be used to strip off advertisements, (5) parsed text, including key words, (6) unparsed text, which could include the whole document, (7) the type of source, including the Internet domain name, (8) date and time of document, (9) any

patent field, such as assignee, inventor name, class, prior art citations, etc., and (10) other user-assigned attributes.

The availability of user-assigned attributes makes the present invention a very powerful tool in which, for example, a sliding scale of file size versus date could be used, or a special weighting for a particular attribute could be used. Furthermore, as described above, an image can be selected to represent a particular document, or a particular set of documents that represent certain search terms from the crafted query.

After the image or text data are linked at the step 550, a decision step 560 determines whether or not the new data should automatically be linked to existing data already residing in one of the thesauri databases. If the answer is YES, then a function step 570 will automatically link data into existing database records on the bulk memory storage device 34. If the answer is NO at decision step 560, then the linking of new data to existing data will be performed manually by the user.

The final result of the crafted query is a thesaurus database that resides on the bulk memory storage device 34, which information preferably is presented in a graphical format. For example, a hyperbolic tree could be used to show various "clusters" of patent documents that are grouped by a certain common attribute. Each "node" of the hyperbolic tree could represent one cluster of patents, and such a node could be linked to other similar nodes that have different clusters of patents with attributes that have some commonality but also certain significant differences, such that these references are placed into different nodes. Another and more preferred method of presenting the database information is a three-dimensional space that can virtually be moved through by user controls to find various groupings of patent documents or other types of technical references that have certain common attributes. Such a three-dimensional space presentation computer program is described in a commonly-assigned United States Patent application titled Method and Apparatus for Interactively Displaying Three-Dimensional Representations of Database Contents, filed on March 8, 1999, and having the serial number 09/_____.

The foregoing description of a preferred embodiment of the invention has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Obvious modifications or variations are possible in light of the above teachings. The embodiment was chosen and described in order to best illustrate the principles of the invention and its practical application to thereby enable one of ordinary skill in the art to best utilize the invention in various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the claims appended hereto.

What is claimed is:

1. A method for searching documents in a database, comprising the steps of: providing a computer having a processing circuit, a memory circuit, a communications circuit, and a video monitor, said computer being configured to communicate with at least one remote database by way of said communications circuit; choosing at least one word as an "initial query" and searching said at least one remote database using said initial query, then displaying an initial result on said video monitor; said method characterized by the steps of:

   iteratively prompting a human user to interactively modify said initial query into a "crafted query" and again searching said at least one remote database using said crafted query, then displaying a modified result on said video monitor; and searching said at least one remote database using a final version of said crafted query and displaying a crafted result on said video monitor.

2. The method as recited in claim 1, wherein said at least one remote database comprises an electronically-accessible bulk memory storage device that contains a plurality of technical reference documents; or wherein said initial query comprises at least one word or phrase; or wherein said crafted query initially comprises said at least one word or phrase in addition to at least one equivalent word or phrase; or wherein said final version of said crafted query comprises a first term and a second term, and said first and second terms are formed into an expression using at least one Boolean operator.

3. The method as recited in claim 2, wherein said plurality of technical reference documents includes patent documents; or wherein said electronically-accessible bulk memory storage device comprises a network server operatively connected to said computer through a local area network; or wherein said electronically-accessible bulk memory storage device comprises an Internet web site having a network server that is operatively connected to said computer through the Internet; or wherein said final version of said crafted query comprises a first term and

equivalent other terms, and wherein said first term and equivalent other terms are formed into an expression using at least one Boolean operator.

4. The method as recited in any above claim, further comprising the step of associating said initial query with at least one image that is displayed on said video monitor in a location proximal to textual information of said initial query.

5. The method as recited in claim 4, further comprising the step of associating said crafted query with at least one image that is displayed on said video monitor in a location proximal to textual information of said crafted query.

6. The method as recited in claim 1, further comprising the step of providing linking criteria that is used to link together a plurality of fields of information from at least one on-line document found in said at least one remote database; or further comprising the step of creating at least one technical thesaurus using said crafted result.

7. The method as recited in claim 6, wherein said linking criteria is used on a sliding scale.

8. A networked computer system used for searching documents in a database, comprising: a computer having a processing circuit, a memory circuit, a communications circuit, and a video monitor, said computer being configured to communicate with at least one remote database over a network by way of said communications circuit; and choose at least one word as an "initial query" and to search said at least one remote database using said initial query, then to display an initial result on said video monitor; said computer system being characterized in that:

said computer iteratively prompts a human user to interactively modify said initial query into a "crafted query" and again to search said at least one remote database using said crafted query, then displays a modified result on said video monitor; and said computer searches said at least one remote database using a

final version of said crafted query, and displays a crafted result on said video monitor.

9. The computer system as recited in claim 8, wherein said at least one remote database comprises an electronically-accessible bulk memory storage device that contains a plurality of technical reference documents; or wherein said final version of said crafted query comprises a first term and a second term, and said first and second terms are formed into an expression using at least one Boolean operator; or wherein said crafted query initially comprises said at least one word or phrase in addition to at least one equivalent word or phrase; or wherein said initial query comprises at least one word or phrase; or wherein said computer is further configured to provide linking criteria that is used to link together a plurality of fields of information from at least one on-line document found in said at least one remote database; or wherein said computer is further configured to create at least one technical thesaurus using said crafted result; or wherein said computer is further configured to associate said initial query with at least one image that is displayed on said video monitor in a location proximal to textual information of said initial query; or wherein said computer is further configured to associate said crafted query with at least one image that is displayed on said video monitor in a location proximal to textual information of said crafted query.

10. The computer system as recited in claim 9, wherein said plurality of technical reference documents includes patent documents; or wherein said electronically-accessible bulk memory storage device comprises a network server operatively connected to said computer through a local area network; or wherein said electronically-accessible bulk memory storage device comprises an Internet web site having a network server that is operatively connected to said computer through the Internet; or wherein said final version of said crafted query comprises a first term and equivalent other terms, and wherein said first term and equivalent other terms are formed into an expression using at least one Boolean operator; or wherein said linking criteria is used on a sliding scale.

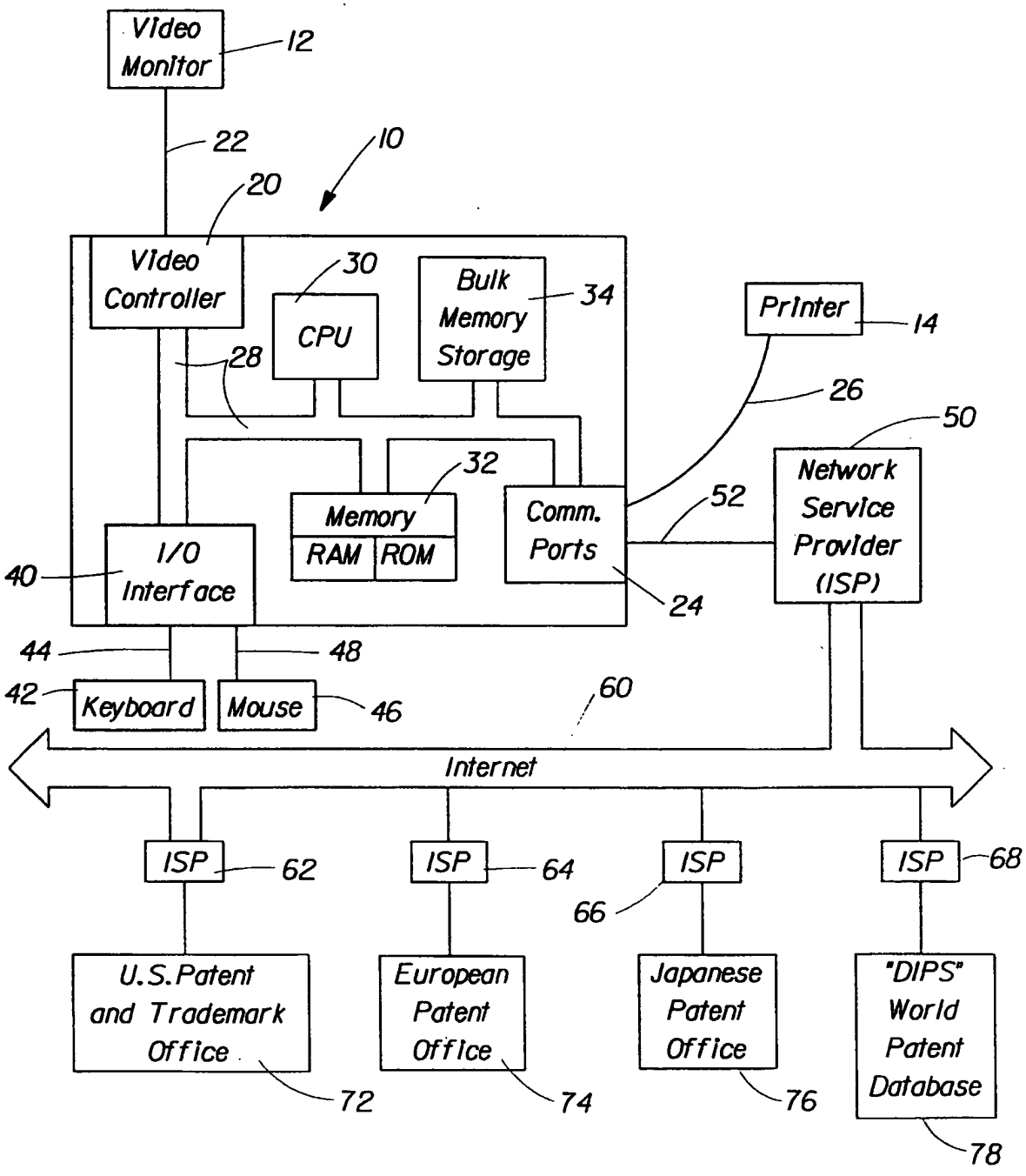

Video Monitor
12
22
10
20
Video Controller
30
CPU
Bulk Memory Storage
34
Printer
14
28
26
50
32
Memory
RAM
ROM
Comm. Ports
52
Network Service Provider (ISP)
40
I/O Interface
24
44
48
42
Keyboard
Mouse
46
60
Internet
ISP
62
ISP
64
ISP
66
ISP
68
U.S. Patent and Trademark Office
72
European Patent Office
74
Japanese Patent Office
76
"DIPS" World Patent Database
78

Fig. 1

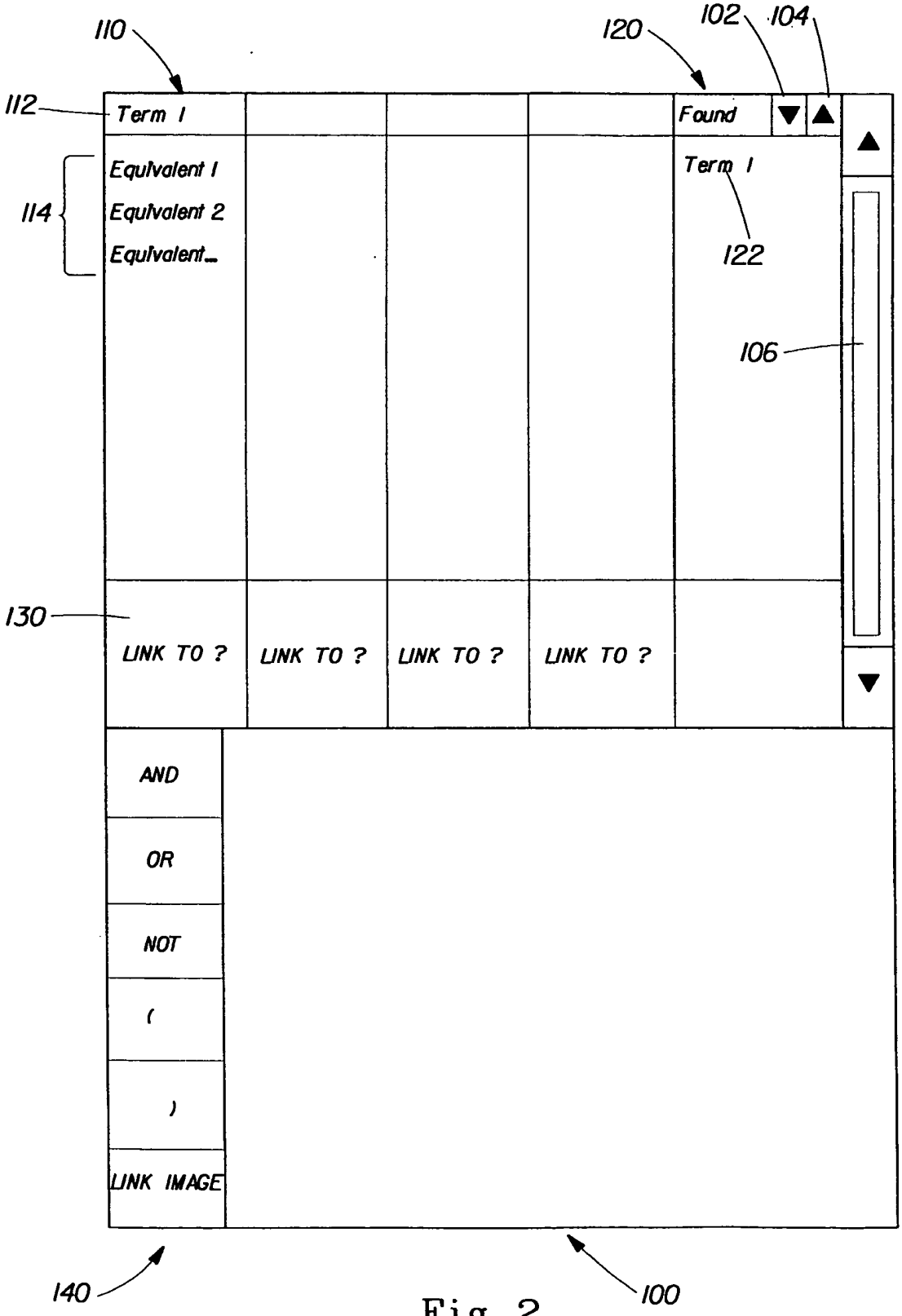

110
120
102
104
112
Term 1
Found
Equivalent 1
Equivalent 2
Equivalent...
114
Term 1
122
106
130
LINK TO ?
LINK TO ?
LINK TO ?
LINK TO ?
AND
OR
NOT
(
)
LINK IMAGE
140
100

Fig. 2

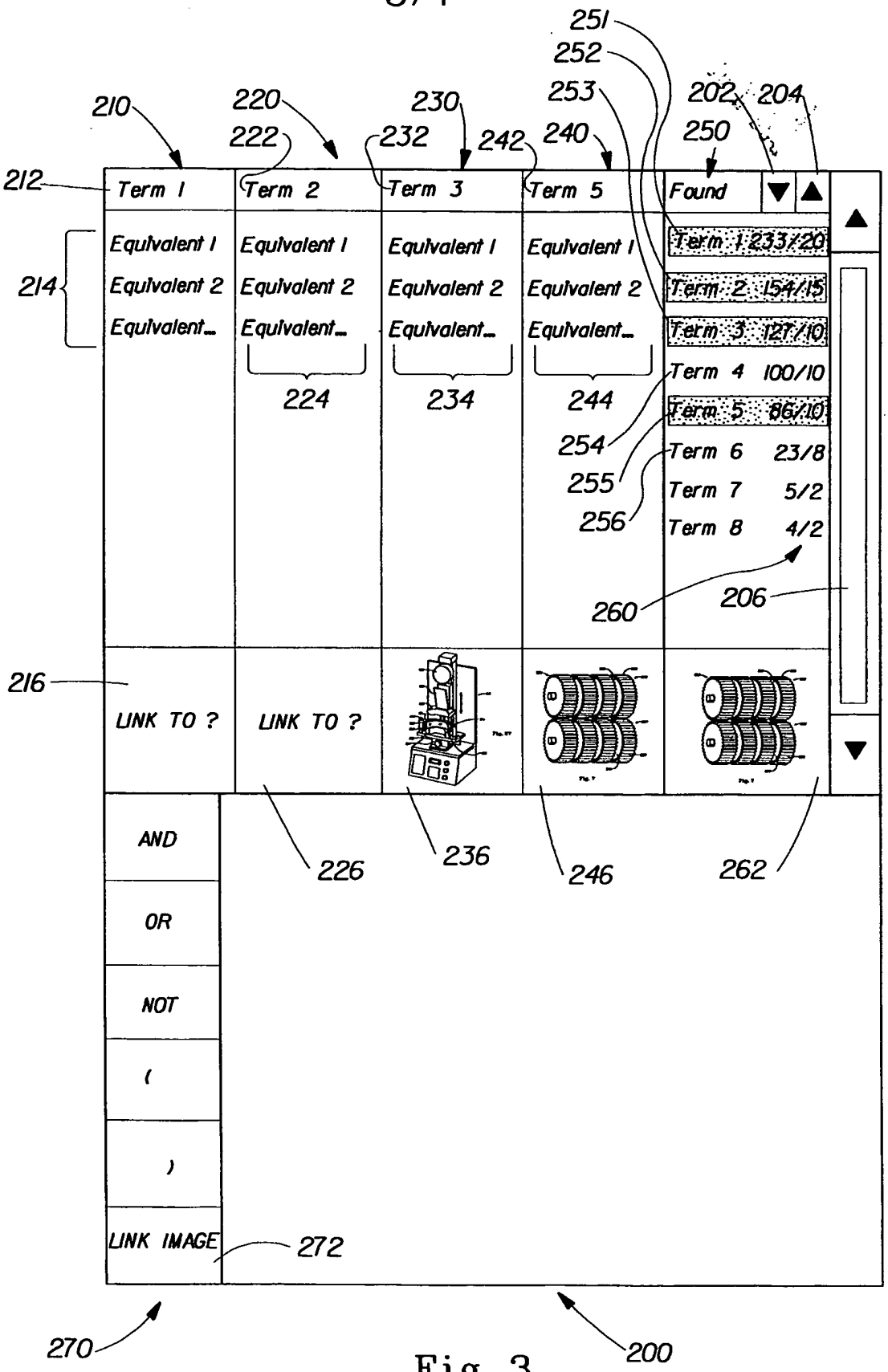

210
220
230
240
250
202
204
222
232
242
212
Term 1
Term 2
Term 3
Term 5
Found
214
Equivalent 1
Equivalent 2
Equivalent...
224
234
244
Term 1 233/20
Term 2 154/15
Term 3 127/10
Term 4 100/10
Term 5 86/10
Term 6 23/8
Term 7 5/2
Term 8 4/2
251
252
253
254
255
256
260
206
216
LINK TO ?
LINK TO ?
226
236
246
262
AND
OR
NOT
(
)
LINK IMAGE
272
270
200

Fig. 3

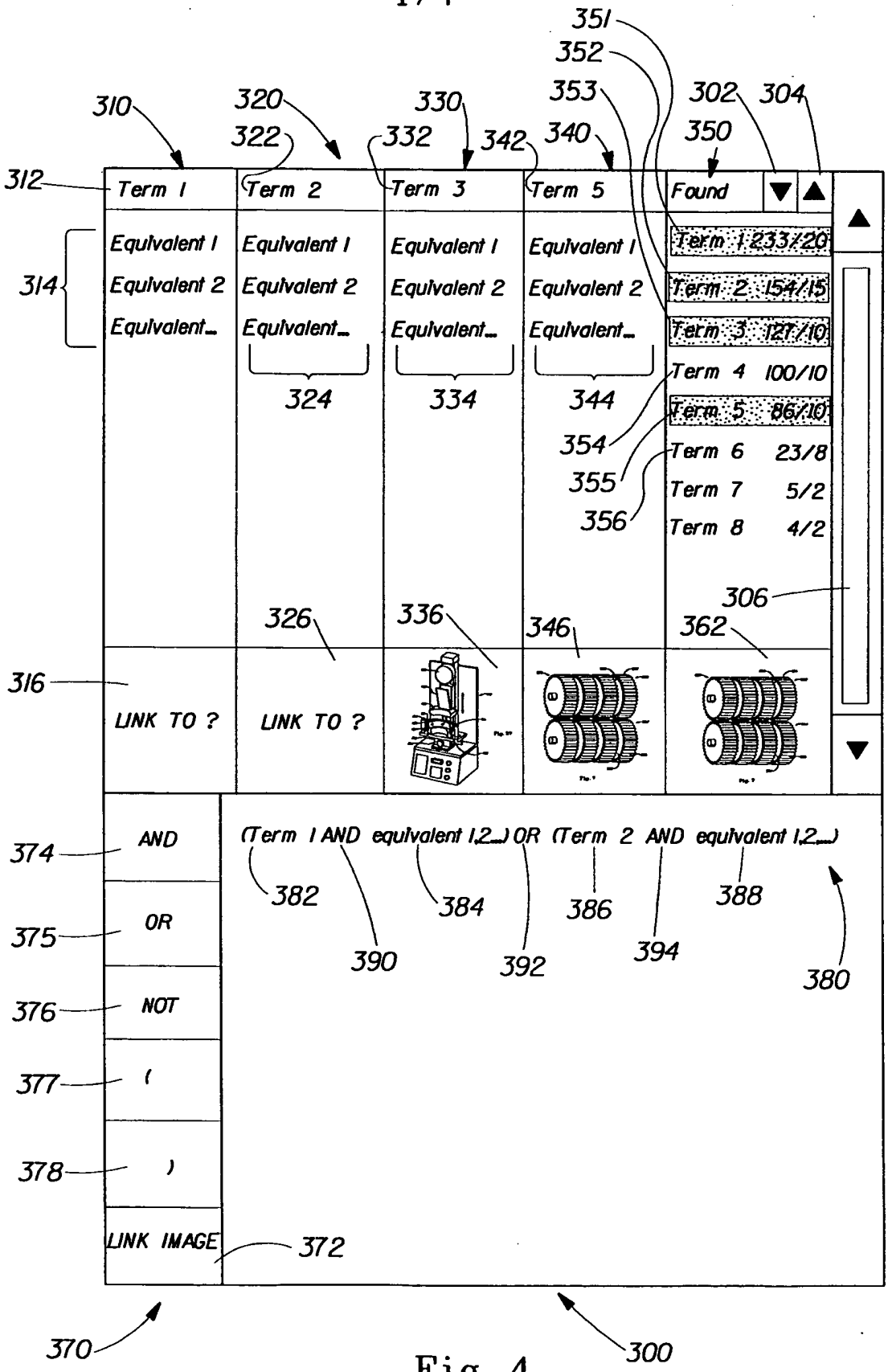

351
352
353
310
320
330
302
304
322
332
342
340
350
312
Term 1
Term 2
Term 3
Term 5
Found
Equivalent 1
Equivalent 2
Equivalent...
Term 1 233/20
Term 2 154/15
Term 3 127/10
Term 4 100/10
Term 5 86/10
Term 6 23/8
Term 7 5/2
Term 8 4/2
314
324
334
344
354
355
356
306
326
336
346
362
316
LINK TO ?
LINK TO ?
374
AND
375
OR
376
NOT
377
(
378
)
LINK IMAGE
(Term 1 AND equivalent 1,2...) OR (Term 2 AND equivalent 1,2...)
382
384
386
388
390
392
394
380
372
370
300

Fig. 4

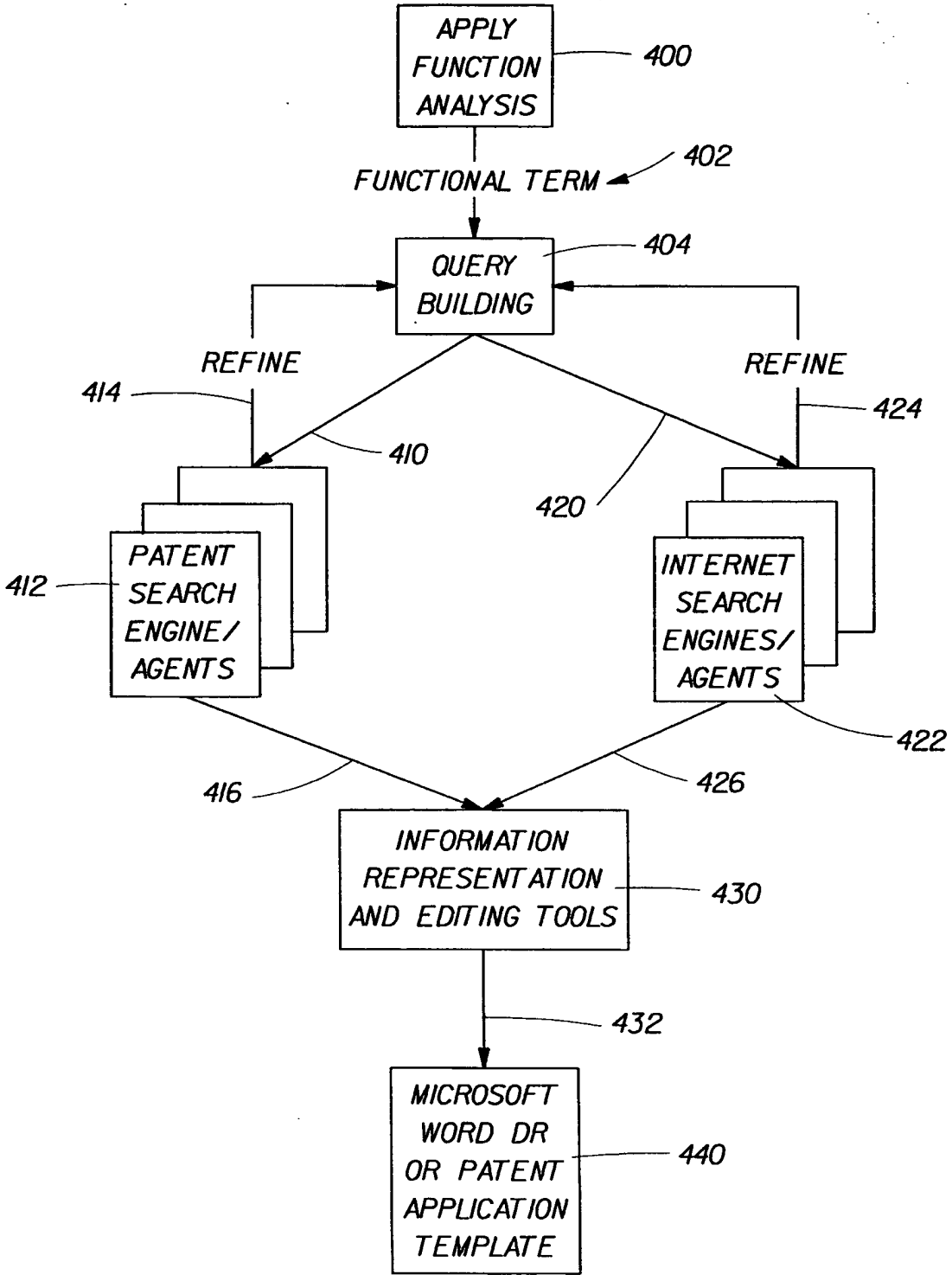

APPLY FUNCTION ANALYSIS
400
FUNCTIONAL TERM
402
QUERY BUILDING
404
REFINE
414
410
REFINE
424
420
PATENT SEARCH ENGINE/ AGENTS
412
INTERNET SEARCH ENGINES/ AGENTS
422
416
426
INFORMATION REPRESENTATION AND EDITING TOOLS
430
432
MICROSOFT WORD DR OR PATENT APPLICATION TEMPLATE
440

Fig. 5

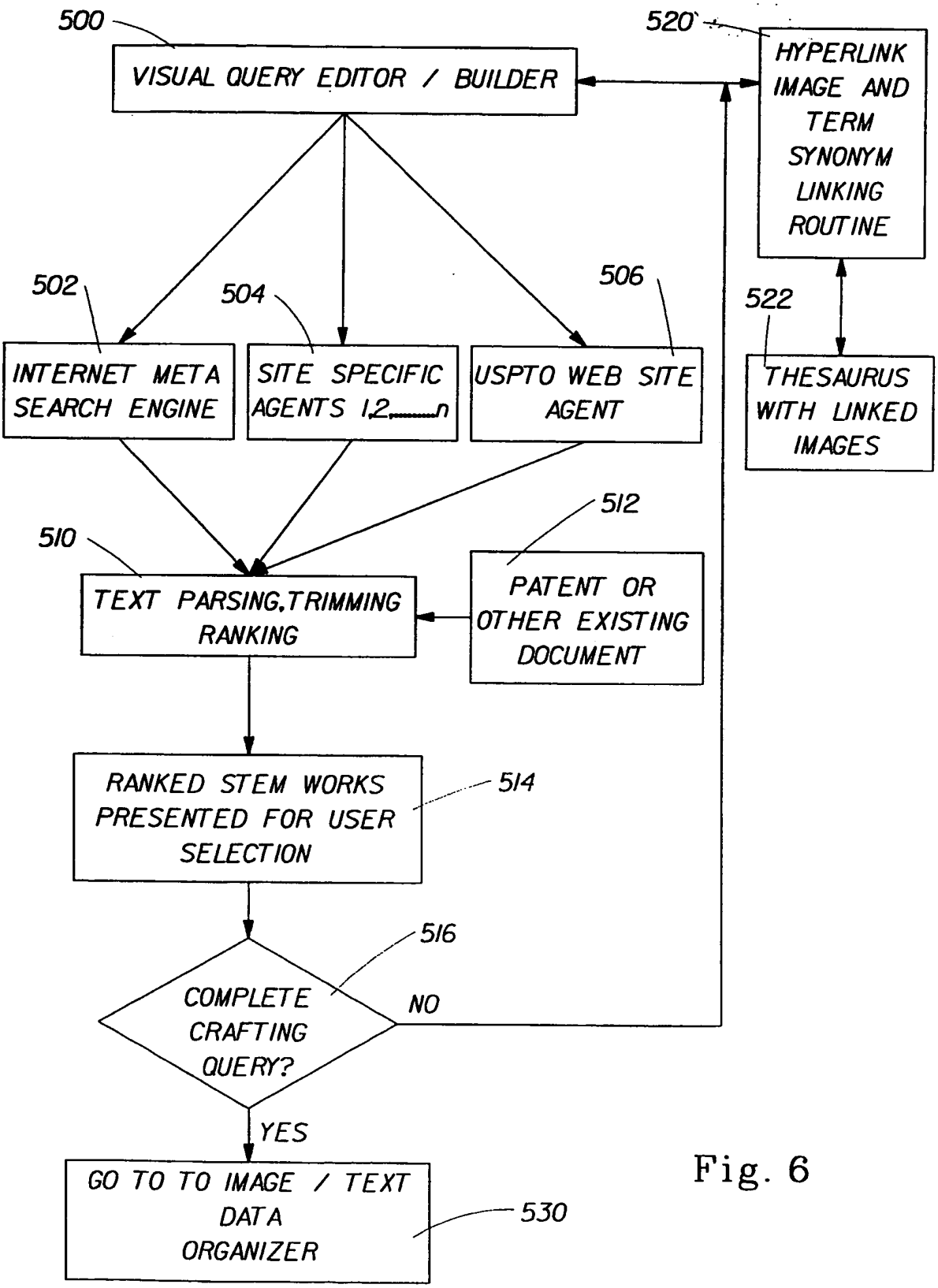

500
VISUAL QUERY EDITOR / BUILDER
520
HYPERLINK IMAGE AND TERM SYNONYM LINKING ROUTINE
502
INTERNET META SEARCH ENGINE
504
SITE SPECIFIC AGENTS 1,2,.....n
506
USPTO WEB SITE AGENT
522
THESAURUS WITH LINKED IMAGES
512
510
TEXT PARSING, TRIMMING RANKING
PATENT OR OTHER EXISTING DOCUMENT
RANKED STEM WORKS PRESENTED FOR USER SELECTION
514
516
COMPLETE CRAFTING QUERY?
NO
YES
GO TO TO IMAGE / TEXT DATA ORGANIZER
530

Fig. 6

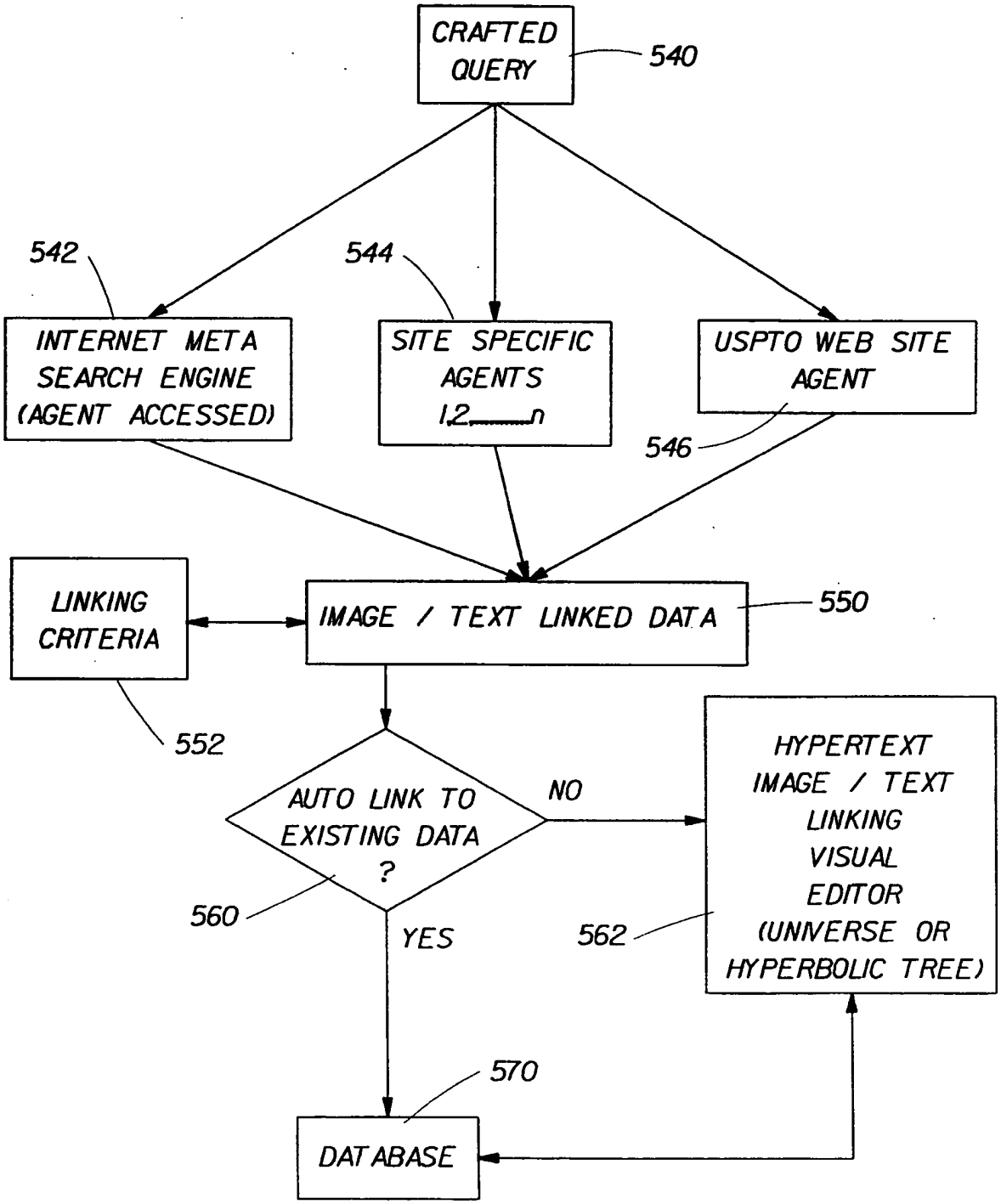

CRAFTED QUERY
540
542
544
INTERNET META SEARCH ENGINE (AGENT ACCESSED)
SITE SPECIFIC AGENTS 1,2,....n
USPTO WEB SITE AGENT
546
LINKING CRITERIA
IMAGE / TEXT LINKED DATA
550
552
AUTO LINK TO EXISTING DATA ?
NO
HYPERTEXT IMAGE / TEXT LINKING VISUAL EDITOR (UNIVERSE OR HYPERBOLIC TREE)
560
YES
562
570
DATABASE

Fig. 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/06072

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :G06F 17/30

US CL :707/2, 3, 4, 5, 10

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 707/2, 3, 4, 5, 10

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EAST

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	LEXIS-NEXIS, Learning Lexis, A Handbook For Mordern Legal Research, LEXIS-NEXIS, Reed Elsevier Inc. 1995, pages 19 and 24.	1-10
Y,P	US 6,006,225 A (BOWMAN et al) 21 December 1999, Abstract, column 6, lines 3-column 14, lines 45.	1-10

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance, the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance, the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search 30 MAY 2000	Date of mailing of the international search report 13 JUN 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer KIM YEN VU James R. Matthews Telephone No. (703) 305-4393

Form PCT/ISA/210 (second sheet) (July 1998)*